



TITLE:

Machinery of Numerical Instability in Conservative Difference Approximations for Compressible Euler Equations (Mathematical Analysis in Fluid and Gas Dynamics)

AUTHOR(S):

Aiso, Hideaki; Takahashi, Tadayasu; Abouziarov,
Moustafa

CITATION:

Aiso, Hideaki ...[et al]. Machinery of Numerical Instability in Conservative Difference Approximations for Compressible Euler Equations (Mathematical Analysis in Fluid and Gas Dynamics). 数理解析研究所講究録 2004, 1353: 178-191

ISSUE DATE:

2004-01

URL:

<http://hdl.handle.net/2433/25154>

RIGHT:

Machinery of Numerical Instability in Conservative Difference Approximations for Compressible Euler Equations.

航空宇宙技術研究所 CFD 技術開発センター

相曾 秀昭 (AISO, Hideaki) 高橋 匡康 (TAKAHASHI, Tadayasu)

Computational Sciences Div., National Aerospace Lab. JAPAN, Jindaiji-Higashi 7-44-1
Chofu TOKYO 185-8522 JAPAN.

航空宇宙技術研究所 CFD 技術開発センター、ニジニノヴゴロド大学力学研究所

アブジアロフ ムスタファ (ABOUZIAROV, Moustafa)

Computational Sciences Div., National Aerospace Lab. JAPAN, Jindaiji-Higashi 7-44-1
Chofu TOKYO 185-8522 JAPAN.

Institute of Mechanics, Nizhni-Novgorod University.

1 Introduction

During almost half a century following the invention of Lax-Friedrichs scheme, which is the first numerical algorithm that gives stable numerical solutions to the compressible Euler equations, various difference schemes have been proposed by many authors. The schemes have been already examined by various test problems and practical examples, and now in the field of numerical computation we have experience somehow enough to obtain numerical results to wide range of problems by choosing some appropriate difference scheme considering the cost and required quality of computation. But any scheme can not be used yet to prove the existence or uniqueness of solution to the compressible Euler equations, which means that any difference scheme is not completely guaranteed in mathematical sense. Even from the viewpoint of practical computation, we often have to make several trials with different schemes to select an appropriate one. In other words, we still have essential problems to be solved in the field of difference schemes.

So called the “numerical carbuncle”^[3]¹ ^[4]² is one of such open problems in the computation. This strange numerical instability often happens and grows around shock wave surface in the numerical simulation of gas flow governed by the compressible Euler equations. While the instability is very small and invisible in the beginning of computation, it may grow as the computation goes on and finally destroy the computation. From the viewpoint of practical treatment to avoid this instability, it is already known that some additional numerical viscosity works well enough to suppress the instability, while it is inevitable to deteriorate the quality of computational result. Therefore it is naturally required to discuss the instability theoretically and to know how much numerical viscosity is the minimum to suppress the growth of instability.

In this article we try to give some mathematical explanation on the machinery that makes the instability, and we show some numerical experiments to verify the explanation.

¹This instability should have been observed even before [3], but it seemed to have been recognized not as a part of property of numerical algorithm but as a kind of quantization error just coming from the digital computation. In [3] it was mentioned for the first time that the instability might have some relation with the essential property of difference scheme or numerical algorithm.

²The article [4], which does not aim so much theoretical discussion, is written from rather comprehensive viewpoint and observes various examples obtained by several different schemes. This is one of the best article to know the situation of research on this issue at the time of writing.

This article is organized as follows. In sections 2 and 3 we review the compressible Euler equation, Godunov method, and the phenomenon of numerical carbuncle. In section 4 the main theorem is given to explain the machinery of growth of numerical carbuncle. In section 5 we verify the formula in main theorem by some numerical examples.

2 Compressible Euler Equations and Godunov Method

We are concerned with difference approximation for the compressible Euler equations in xy -space;

$$U_t + F_x + G_y = 0, \quad -\infty < x < \infty, -\infty < y < \infty, 0 < t < \infty, \quad (1)$$

$$U = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ e \end{bmatrix}, F = F(U) = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(e + p) \end{bmatrix}, G = G(U) = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(e + p) \end{bmatrix},$$

where ρ, u, v, p, e are density, velocity in x -direction, velocity in y -direction, pressure, total energy per unit volume of the fluid, respectively. The total energy e is determined by the equation of state;

$$e = \frac{p}{\gamma - 1} + \frac{1}{2}\rho(u^2 + v^2) \quad (2)$$

with the adiabatic constant γ . U is called the vector of conservative variables and F, G are called flux functions in x - and y -direction, respectively. By V , we mean the vector of so called primitive variables;

$$V = \begin{bmatrix} \rho \\ u \\ v \\ p \end{bmatrix}. \quad (3)$$

We employ Godunov method [1] to discretize the problem (1). Godunov method is one of difference approximation based on the concept of finite volume and Riemann problem. It is given as follows.

The xy -space $(-\infty, \infty) \times (-\infty, \infty)$ is divided into the set $\{I_{i,j}\}_{i,j:\text{integer}}$ of finite volumes $I_{i,j} = \left((i - \frac{1}{2})\Delta x, (i + \frac{1}{2})\Delta x\right) \times \left((j - \frac{1}{2})\Delta y, (j + \frac{1}{2})\Delta y\right)$, where Δx and Δy are spatial difference increments in x - and y -directions, respectively. The node $(i\Delta x, j\Delta y)$ represents each finite volume $I_{i,j}$. The discretization of time t is given by $0 = t^0 < t^1 < \dots < t^n < t^{n+1} < \dots$, where the temporal increment Δt^n is determined by $\Delta t^n = t^{n+1} - t^n$. The approximation of values ρ, u, v, p, e over each $I_{i,j}$ (or at the node $(i\Delta x, j\Delta y)$) at the time $t = t^n$ is written by $\rho_{i,j}^n, u_{i,j}^n, v_{i,j}^n, p_{i,j}^n, e_{i,j}^n$, respectively. Throughout the discussion we assume the relation

$$e_i^n = \frac{p_i^n}{\gamma - 1} + \frac{1}{2}\rho_i^n \left((u_i^n)^2 + (v_i^n)^2 \right) \quad (4)$$

for all n, i, j to be consistent with (2). The discretized temporal evolution of approximate values of conservative variables

$$U_{i,j}^n = \begin{bmatrix} \rho_{i,j}^n \\ \rho_{i,j}^n u_{i,j}^n \\ \rho_{i,j}^n v_{i,j}^n \\ e_{i,j}^n \end{bmatrix}$$

is given by the difference scheme

$$U_{i,j}^{n+1} = U_{i,j}^n - \frac{\Delta t^n}{\Delta x} \left\{ \bar{F}_{i+\frac{1}{2},j}^n - \bar{F}_{i-\frac{1}{2},j}^n \right\} - \frac{\Delta t^n}{\Delta y} \left\{ \bar{G}_{i,j+\frac{1}{2}}^n - \bar{G}_{i,j-\frac{1}{2}}^n \right\}, \quad (5)$$

where the numerical fluxes $\bar{F}_{i+\frac{1}{2},j}^n$, $\bar{G}_{i,j+\frac{1}{2}}^n$ in x - and y - directions, respectively are given as follows.

First we assume the Riemann problem

$$\begin{cases} U_t + F_x = 0 \\ U(x, 0) = \begin{cases} U_{i,j}^n, & x < 0 \\ U_{i+1,j}^n, & x > 0. \end{cases} \end{cases} \quad (6)$$

given by the states of neighboring finite volumes $I_{i,j}$ and $I_{i+1,j}$. Using the exact solution to (6), which is self similar; $U = U(x, t) = U(x/t; U_{i,j}^n, U_{i+1,j}^n)$, we determine the numerical flux $\bar{F}_{i+\frac{1}{2},j}^n$ by

$$\bar{F}_{i+\frac{1}{2},j}^n = F(\bar{U}_{i+\frac{1}{2},j}^n), \quad (7)$$

where $\bar{U}_{i+\frac{1}{2},j}^n$ is given by

$$\bar{U}_{i+\frac{1}{2},j}^n = U(0; U_{i,j}^n, U_{i+1,j}^n).$$

$\bar{U}_{i+\frac{1}{2},j}^n$ is regarded as a kind of virtual state assumed at the contact $\{(i + \frac{1}{2})\Delta x\} \times ((j - \frac{1}{2})\Delta y, (j + \frac{1}{2})\Delta y)$ between $I_{i,j}$ and $I_{i+1,j}$ to determine the numerical flux $\bar{F}_{i+\frac{1}{2},j}^n$. The numerical flux $\bar{G}_{i,j+\frac{1}{2}}^n$ is obtained in a similar manner. From the exact solution $U = U(y, t) = U(y/t; U_{i,j}^n, U_{i,j+1}^n)$ to the Riemann problem

$$\begin{cases} U_t + G_y = 0 \\ U(y, 0) = \begin{cases} U_{i,j}^n, & y < 0 \\ U_{i,j+1}^n, & y > 0, \end{cases} \end{cases} \quad (8)$$

we determine

$$\bar{G}_{i,j+\frac{1}{2}}^n = G(\bar{U}_{i,j+\frac{1}{2}}^n), \quad (9)$$

where

$$\bar{U}_{i,j+\frac{1}{2}}^n = U(0; U_{i,j}^n, U_{i,j+1}^n).$$

3 Numerical Carbuncle

"Numerical carbuncle" (or "carbuncle instability", "carbuncle phenomenon") is numerical instability that may happen when the numerical computation includes a strong shock wave.

Sometimes it happens a misunderstanding that the computation simulates so called physical carbuncle, a physical phenomenon that the surface of a shock wave in dusty air is fluctuated by dust particles. But in the case of numerical carbuncle the assumed governing equation is just the compressible Euler equations where the situation of dusty air is never taken into account. Therefore it is reasonable to think that the numerical algorithm to approximate the compressible Euler equation might include some machinery that makes the numerical instability.

From the experience of numerical computation, the following empirical facts are observed.

1. The instability occurs not in one-dimensional numerical computation but in more than two dimensional cases. Even in the case of one dimensional physical phenomenon to be simulated, the instability may happen if the numerical computation is done in two dimension. (The example discussed in this article is one of the most typical cases. A progressing planar shock wave is a one dimensional phenomenon. But, if the computation for the problem is done in two dimension, the instability may occur.)
2. When the shock surface is oblique enough to any of axes of the discretization mesh, the instability does not occur.
3. The instability seems to be initiated round off error that is from quantization, *i.e.* essential error in digital computation.
4. Once the instability is observed, the growth rate of instability seems to be exponential with respect to the step number n for the temporal discretization.
5. The increase of numerical viscosity is useful to suppress the instability. The numerical viscosity in the direction along shock surface works more effective.
6. When the instability is small enough, the perturbation of each value (mass, momentum in each direction, or pressure) seems to have "odd-even" property [3], *i.e.* the perturbations of each value at any neighboring computing nodes (finite volumes) have the opposite signs.

The experiences above implies the followings.

1. If the distribution of variables in discretized model are completely one dimensional at some time step n (*i.e.* $U_{i,j}^n$ does not depend on j), there is no reason that the exact computation of algorithm determined by (5)-(9) makes any loss of one dimensional property at the next time step $n + 1$. Therefore, some error derived from digital computation should initialize the instability (or make some numerical perturbation that is grown up to the instability).
2. Error from digital computation is understood to have pseudo-stochastic property. It implies that the carbuncle may grow almost in the order of \sqrt{n} if it is only because of this error. But the observed fact is different. We may expect that the numerical algorithm includes some machinery to amplify the error at each step of temporal evolution.

The discussion on the occurrence of error in digital computation is not so easy because it depends on the hardware architecture and operationing system etc. of each computer. Therefore we assume that some small numerical error is already given. Then we discuss how the error propagates in the procedure of temporal evolution given by Godunov method (5)-(9).

4 Analysis of Noise Propagation

In this section we assume that some small perturbation is given to numerical values $*_{i,j}^n$ at the time step n and we then analyze how the small perturbation propagates in the procedure of discretized temporal evolution from the time step n to $n + 1$.

Our analysis is done to Godunov method given by (5)-(9). We have some additional assumption.

Assumption 1 There are some $\rho_i^n, u_i^n, p_i^n, e_i^n$ (for all n, i) and $\hat{\rho}_{i,j}^n, \hat{u}_{i,j}^n, \hat{v}_{i,j}^n, \hat{p}_{i,j}^n, \hat{e}_{i,j}^n$ (for all n, i, j) that satisfy

$$\begin{cases} \rho_{i,j}^n = \rho_i^n + \hat{\rho}_{i,j}^n \\ u_{i,j}^n = u_i^n + \hat{u}_{i,j}^n \\ v_{i,j}^n = \hat{v}_{i,j}^n \\ p_{i,j}^n = p_i^n + \hat{p}_{i,j}^n \\ e_{i,j}^n = e_i^n + \hat{e}_{i,j}^n, \end{cases} \quad (10)$$

where each of $\hat{\rho}_{i,j}^n, \hat{u}_{i,j}^n, \hat{v}_{i,j}^n, \hat{p}_{i,j}^n, \hat{e}_{i,j}^n$ is small enough. We chose some representative δ of the order of them;

$$|\hat{\rho}_{i,j}^n|, |\hat{u}_{i,j}^n|, |\hat{v}_{i,j}^n|, |\hat{p}_{i,j}^n|, |\hat{e}_{i,j}^n| \leq O(\delta) \quad (11)$$

We also assume that the relation of discretized temporal evolution of Godunov method (5)-(9) should be satisfied between $\{U_{i,j}^n\}_{i,j}$ and $\{U_{i,j}^{n+1}\}_{i,j}$ when $\hat{\rho}_{i,j}^n = \hat{u}_{i,j}^n = \hat{v}_{i,j}^n = \hat{p}_{i,j}^n = \hat{e}_{i,j}^n = 0$ for all n, i, j .

Under assumption 1 we employ the following notation.

$$U_i^n = \begin{bmatrix} \rho_i^n \\ \rho_i^n u_i^n \\ 0 \\ e_i^n \end{bmatrix}, V_i^n = \begin{bmatrix} \rho_i^n \\ u_i^n \\ 0 \\ p_i^n \end{bmatrix}, \hat{U}_{i,j}^n = U_{i,j}^n - U_i^n, \hat{V}_{i,j}^n = \begin{bmatrix} \hat{\rho}_{i,j}^n \\ \hat{u}_{i,j}^n \\ \hat{v}_{i,j}^n \\ \hat{p}_{i,j}^n \end{bmatrix}. \quad (12)$$

Assumption 2 Each u_i^n satisfies

$$u_i^n \gg c_i^n, \quad (13)$$

where $c_i^n = \sqrt{\frac{\gamma p_i^n}{\rho_i^n}}$.

Assumption 3. Each Δt^n should satisfy

$$\left(|u_{i,j}^n| + |c_{i,j}^n| \right) \frac{\Delta t^n}{\Delta x}, \left(|v_{i,j}^n| + |c_{i,j}^n| \right) \frac{\Delta t^n}{\Delta y} \leq C, \quad (14)$$

where C is a positive constant less than 1.

Assumption 1 means that the situation is almost one dimensional and the essential direction of phenomenon is x -direction. Assumption 2 means the complete upwindness in x -direction. In other words, we assume it so that lemma 3 below could be applied, i.e. the characteristic at each finite volume or those caused by Riemann problem (6) should have positive velocity. Assumption 3 is a usual CFL-condition, a basic stability condition for discretized temporal evolution.

Now we analyze the discretized temporal evolution of perturbations, i.e. the relation between $\hat{\rho}_{i,j}^n, \hat{u}_{i,j}^n, \hat{v}_{i,j}^n, \hat{p}_{i,j}^n, \hat{e}_{i,j}^n$ and $\hat{\rho}_{i,j}^{n+1}, \hat{u}_{i,j}^{n+1}, \hat{v}_{i,j}^{n+1}, \hat{p}_{i,j}^{n+1}, \hat{e}_{i,j}^{n+1}$. We obtain the following theorem especially for $v_{i,j}^{n+1}$.

Theorem 1 *We assume Godunov method (5)-(9) and assumptions 1-3. Then we obtain the following formula.*

$$\begin{aligned} \hat{v}_{i,j}^{n+1} = & \hat{v}_{i,j}^n - \frac{\Delta t^n}{\Delta x} \frac{\rho_{i-1}^n}{\rho_i^{n+1}} \cdot u_{i-1}^n (\hat{v}_{i,j}^n - \hat{v}_{i-1,j}^n) \\ & - \frac{\Delta t^n}{\Delta y} \left\{ \frac{1}{2\rho_i^{n+1}} (\hat{p}_{i,j+1}^n - \hat{p}_{i,j-1}^n) - \frac{\rho_i^n}{\rho_i^{n+1}} \frac{c_i^n}{2} (\hat{v}_{i,j-1}^n - 2\hat{v}_{i,j}^n + \hat{v}_{i,j+1}^n) \right\} \\ & + o(\delta). \end{aligned} \quad (15)$$

This is the main theorem and formula to imply the machinery that the discretized model of Godunov method amplifies the numerical carbuncle. While we show some verification of the formula (15) via some examples of numerical simulation in the next section, here we have some theoretical discussion on the formula. When the numerical data, especially the density, is smooth (*i.e.* it does not contain so big a gradient in x -direction), there is no amplification machinery. But, once there is a big gradient of the density like a shock wave, the factor $\frac{\rho_{i-1}^n}{\rho_i^{n+1}}$ or $\frac{\rho_i^n}{\rho_i^{n+1}}$ would be larger enough than 1 and it implies that the perturbation $\{\hat{v}_{i,j}^n\}_{i,j}$ might be amplified in the temporal evolution from the time step n to $n+1$.

It is natural to assume the odd even property of perturbations according to the observation given by [3]. In this case we have the following corollary.

Corollary 2 *When the perturbation $\{\hat{v}_{i,j}^n\}_{i,j}$ at the time step n satisfies the odd-even property;*

$$\hat{v}_{i,j}^n = (-1)^{i+j} \hat{v}^n, \quad \hat{p}_{i,j}^n = (-1)^{i+j} \hat{p}^n \quad (16)$$

then the formula (15) in theorem 1 is written in the following form.

$$\hat{v}_{i,j}^{n+1} = (-1)^{i+j} \hat{v}^n \left\{ 1 - 2 \left(\frac{\Delta t^n}{\Delta x} \frac{\rho_{i-1}^n}{\rho_i^{n+1}} u_{i-1}^n + \frac{\Delta t^n}{\Delta y} \frac{\rho_i^n}{\rho_i^{n+1}} c_i^n \right) \right\} + o(\delta). \quad (17)$$

The proof is easily given just by substituting the assumption (16) into the formula (15).

We remember that also the usual CFL condition for the stability can be derived from a similar manner to the corollary above, and we understand that the formula (17) gives a stability condition, which requires that the condition

$$\frac{\Delta t^n}{\Delta x} \frac{\rho_{i-1}^n}{\rho_i^{n+1}} u_{i-1}^n + \frac{\Delta t^n}{\Delta y} \frac{\rho_i^n}{\rho_i^{n+1}} c_i^n \leq 1 \quad (18)$$

should be satisfied. We easily have the following observation on the relation among the condition (18), the usual CFL condition (14) and the smoothness of numerical data for density ρ .

Observation *Under the situation assumed here, the satisfaction of condition (18) depends on the smoothness of the numerical data for density ρ .*

1. *We suppose that the numerical data for density ρ is smooth enough, *i.e.* the ratios $\frac{\rho_{i-1}^n}{\rho_i^{n+1}}$ and $\frac{\rho_i^n}{\rho_i^{n+1}}$ are near enough to 1. Then the condition (18) can be guaranteed by the usual CFL condition (14) of assumption 3.*

2. We suppose that the numerical data for density ρ has a gradient big enough.

Then either of $\frac{\rho_{i-1}^n}{\rho_i^{n+1}}$ or $\frac{\rho_i^n}{\rho_i^{n+1}}$ may be larger enough than 1. In result, the stability condition (18) might be violated even though the usual CFL condition (14) of assumption 3 is satisfied.

Therefore, if a big gradient is included in the numerical data, the usual CFL condition (14) is not enough to guarantee the stability of numerical calculation.

From the discussion above we may insist that the machinery represented by the formula (15) in theorem 1 is a main reason of the amplification of perturbations in the numerical carbuncle. In the next section we demonstrates some numerical examples to verify it.

The remaining part of section is devoted to the proof of theorem 1. First we analyze the effect of perturbation $\hat{V}_{i,j}^n$ on each of the flux differences $\bar{F}_{i+\frac{1}{2},j}^n - \bar{F}_{i-\frac{1}{2},j}^n$ and $\bar{G}_{i,j+\frac{1}{2}}^n - \bar{G}_{i,j-\frac{1}{2}}^n$.

From the assumption 2, we can use the following lemma.

Lemma 3 If $u_{i,j}^n$ and $u_{i+1,j}^n$ are larger enough than $c_{i,j}^n$ and $c_{i+1,j}^n$, respectively, the following holds for the numerical flux $\bar{F}_{i+\frac{1}{2},j}^n$.

$$\bar{F}_{i+\frac{1}{2},j}^n = F(U_{i,j}^n) = \begin{bmatrix} \rho_{i,j}^n u_{i,j}^n \\ \rho_{i,j}^n (u_{i,j}^n)^2 + p_{i,j}^n \\ \rho_{i,j}^n u_{i,j}^n v_{i,j}^n \\ u_{i,j}^n (e_{i,j}^n + p_{i,j}^n) \end{bmatrix}. \quad (19)$$

The proof is easily obtained from the basic property of Riemann problem (6) that $\bar{U}_{i+\frac{1}{2},j}^n = U_{i,j}^n$ because of the upwindness. (For example, see [1, 2].)

To discuss $\bar{G}_{i,j+\frac{1}{2}}^n - \bar{G}_{i,j-\frac{1}{2}}^n$ we use linearization. Assumption 1 means that the partial differential equation

$$U_t + G(U)_y = 0 \quad (20)$$

that arises in the Riemann problem (8) is near enough to the linear problem

$$U_t + A_i^n U_y = 0, \quad (21)$$

where

$$A_i^n = A|_{U=U_i^n},$$

and A is Jacobi matrix of the flux function $G(U)$;

$$A = \frac{\partial G}{\partial U}.$$

It is also noted that the Jacobi matrix A is diagonalizable with the eigenvalues $v-c, v, v+c$. Throughout the discussion, let $*_i^n$ and $*_{i,j}^n$ mean $*|_{U=U_i^n}$ and $*|_{U=U_{i,j}^n}$, respectively.

Lemma 4 For the difference $\bar{G}_{i,j+\frac{1}{2}}^n - \bar{G}_{i,j-\frac{1}{2}}^n$ of numerical flux in y -direction, the following approximation holds.

$$\begin{aligned} & \bar{G}_{i,j+\frac{1}{2}}^n - \bar{G}_{i,j-\frac{1}{2}}^n \\ &= \frac{1}{2} A_i^n (U_{i,j+1}^n - U_{i,j-1}^n) - \frac{1}{2} |A_i^n| (U_{i,j-1}^n - 2U_{i,j}^n + U_{i,j+1}^n) + o(\delta), \end{aligned} \quad (22)$$

where the matrix $|A|$ is given by

$$|A| = P|\Lambda|P^{-1} = P \begin{bmatrix} |\lambda_1| & 0 & 0 & 0 \\ 0 & |\lambda_2| & 0 & 0 \\ 0 & 0 & |\lambda_3| & 0 \\ 0 & 0 & 0 & |\lambda_4| \end{bmatrix} P^{-1} \quad (23)$$

using any diagonalization

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix} = P^{-1}AP \quad (24)$$

of the matrix A . It should be noted that $|A|$ is uniquely determined.

Proof:

We consider the following Riemann problem (25), which is the Riemann problem (8) with the PDE replaced by the linearized one (21),

$$\begin{cases} U_t + A_i^n U_y = 0 \\ U(y, 0) = \begin{cases} U_{i,j}^n, & y < 0, \\ U_{i,j+1}^n, & y > 0 \end{cases} \end{cases} \quad (25)$$

and the exact solution $U = U(y, t) = U_{\text{lin}}(y/t; U_{i,j}^n, U_{i,j+1}^n)$ to the problem (25). The linear problem (25) is solved by characteristic decomposition into four scalar equations in characteristic variables. (See basic text books, for example, [2].)

From a matrix A we determine the matrix $\text{sgn}(A)$ by

$$\text{sgn}(A) = P \begin{bmatrix} \text{sgn}(\lambda_1) & 0 & 0 & 0 \\ 0 & \text{sgn}(\lambda_2) & 0 & 0 \\ 0 & 0 & \text{sgn}(\lambda_3) & 0 \\ 0 & 0 & 0 & \text{sgn}(\lambda_4) \end{bmatrix} P^{-1}, \quad (26)$$

using the diagonalization of A ;

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}$$

and the function $\text{sgn}(s)$ (s : any real number);

$$\text{sgn}(s) = \begin{cases} -1, & s < 0 \\ 0, & s = 0 \\ 1, & s > 0. \end{cases}$$

Then we can write the solution U_{lin} in the form

$$U_{\text{lin}}(\xi; U_{i,j}^n, U_{i,j+1}^n) = \frac{1}{2} (U_{i,j}^n + U_{i,j+1}^n) - \frac{1}{2} \text{sgn}(A_i^n - \xi I) (U_{i,j+1}^n - U_{i,j}^n), \quad (27)$$

where I is the unit matrix. We obtain

$$U_{\text{lin}}(0; U_{i,j}^n, U_{i,j+1}^n) = \frac{1}{2} (U_{i,j}^n + U_{i,j+1}^n) - \frac{1}{2} \text{sgn}(A_i^n) (U_{i,j+1}^n - U_{i,j}^n) \quad (28)$$

especially in the case $\xi = 0$.

Then we obtain the following approximation.

$$\begin{aligned} \bar{G}_{i,j+\frac{1}{2}}^n - \bar{G}_{i,j-\frac{1}{2}}^n &= G(\bar{U}_{i,j+\frac{1}{2}}^n) - G(\bar{U}_{i,j-\frac{1}{2}}^n) \\ &= A_i^n (\bar{U}_{i,j+\frac{1}{2}}^n - \bar{U}_{i,j-\frac{1}{2}}^n) + o(\delta) \\ &= A_i^n (U_{\text{lin}}(0; U_{i,j}^n, U_{i,j+1}^n) - U_{\text{lin}}(0; U_{i,j-1}^n, U_{i,j}^n)) + o(\delta) \\ &= \frac{1}{2} A_i^n (U_{i,j+1}^n - U_{i,j-1}^n) - \frac{1}{2} |A_i^n| (U_{i,j-1}^n - 2U_{i,j}^n + U_{i,j+1}^n) + o(\delta), \end{aligned} \quad (29)$$

Here we mention that the relation $A \cdot \text{sgn}(A) = |A|$ is used.

This completes the proof.

Now we observe A and $|A|$. For simplicity, we rewrite the PDE (20) in the following non-conservative form using the primitive variables ρ, u, v, p .

$$\begin{cases} \rho_t + v\rho_y + \rho v_y = 0 \\ u_t + vu_y = 0 \\ v_t + vv_y + \frac{p_y}{\rho} = 0 \\ p_t + \gamma p v_y + v p_y = 0 \end{cases} \quad (30)$$

that are equivalent to

$$V_t + B V_y = 0, \text{ where } V = \begin{bmatrix} \rho \\ u \\ v \\ p \end{bmatrix}, B = \begin{bmatrix} v & 0 & \rho & 0 \\ 0 & v & 0 & 0 \\ 0 & 0 & v & \frac{1}{\rho} \\ 0 & 0 & \gamma p & v \end{bmatrix}. \quad (31)$$

We observe the following facts easily.

Lemma 5 *Between U and V , hold the followings.*

$$\frac{\partial U}{\partial V} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ u & \rho & 0 & 0 \\ v & 0 & \rho & 0 \\ \frac{u^2+v^2}{2} & \rho u & \rho v & \frac{1}{\gamma-1} \end{bmatrix}, \quad (32)$$

and

$$\frac{\partial V}{\partial U} = \left(\frac{\partial U}{\partial V} \right)^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{u}{\rho} & \frac{1}{\rho} & 0 & 0 \\ -\frac{v}{\rho} & 0 & \frac{1}{\rho} & 0 \\ \frac{\gamma-1}{2}(u^2+v^2) & -(\gamma-1)u & -(\gamma-1)v & \gamma-1 \end{bmatrix}. \quad (33)$$

Lemma 6 *The matrix B is diagonalized as follows.*

$$Q^{-1} B Q = \Lambda = \begin{bmatrix} v-c & 0 & 0 & 0 \\ 0 & v & 0 & 0 \\ 0 & 0 & v & 0 \\ 0 & 0 & 0 & v+c \end{bmatrix}, \quad (34)$$

where

$$Q = \begin{bmatrix} \rho & \rho & \rho & \rho \\ 0 & c & -c & 0 \\ -c & 0 & 0 & c \\ \gamma p & 0 & 0 & \gamma p \end{bmatrix} \quad \text{and} \quad Q^{-1} = \begin{bmatrix} 0 & 0 & -\frac{1}{2c} & \frac{1}{2\gamma p} \\ \frac{1}{2\rho} & \frac{1}{2c} & 0 & -\frac{1}{2\gamma p} \\ \frac{1}{2\rho} & -\frac{1}{2c} & 0 & -\frac{1}{2\gamma p} \\ 0 & 0 & \frac{1}{2c} & \frac{1}{2\gamma p} \end{bmatrix}. \quad (35)$$

By the relation $A = \left(\frac{\partial U}{\partial V}\right) B \left(\frac{\partial V}{\partial U}\right)$, which is easily observed, we also obtain the following lemma.

Lemma 7 The matrix $A = \frac{\partial G}{\partial U}$ is diagonalized in the following manner.

$$\left[\left(\frac{\partial U}{\partial V}\right) Q\right]^{-1} A \left[\left(\frac{\partial U}{\partial V}\right) Q\right] = \Lambda = \begin{bmatrix} v - c & 0 & 0 & 0 \\ 0 & v & 0 & 0 \\ 0 & 0 & v & 0 \\ 0 & 0 & 0 & v + c \end{bmatrix}. \quad (36)$$

Using lemmas 5-7, we proceed the calculation from (29);

$$\begin{aligned} & \bar{G}_{i,j+\frac{1}{2}}^n - \bar{G}_{i,j-\frac{1}{2}}^n \\ &= G(\bar{U}_{i,j+\frac{1}{2}}^n) - G(\bar{U}_{i,j-\frac{1}{2}}^n) \\ &= \frac{1}{2} \left(\frac{\partial U}{\partial V}\right)_i^n Q_i^n \Lambda_i^n (Q_i^n)^{-1} \left(\frac{\partial V}{\partial U}\right)_i^n (U_{i,j+1}^n - U_{i,j-1}^n) \\ &\quad - \frac{1}{2} \left(\frac{\partial U}{\partial V}\right)_i^n Q_i^n |\Lambda_i^n| (Q_i^n)^{-1} \left(\frac{\partial V}{\partial U}\right)_i^n (U_{i,j-1}^n - 2U_{i,j}^n + U_{i,j+1}^n) + o(\delta) \\ &= \frac{1}{2} \left(\frac{\partial U}{\partial V}\right)_i^n \left[Q_i^n \Lambda_i^n (Q_i^n)^{-1} (\hat{V}_{i,j+1}^n - \hat{V}_{i,j-1}^n) \right. \\ &\quad \left. - Q_i^n |\Lambda_i^n| (Q_i^n)^{-1} (\hat{V}_{i,j+1}^n - 2\hat{V}_{i,j}^n + \hat{V}_{i,j-1}^n) \right] + o(\delta) \\ &= \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & 0 \\ u_i^n & \rho_i^n & 0 & 0 \\ 0 & 0 & \rho_i^n & 0 \\ \frac{(u_i^n)^2}{2} & \rho_i^n u_i^n & 0 & \frac{1}{\gamma-1} \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & \rho_i^n & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\rho_i^n} \\ 0 & 0 & \gamma p_i^n & 0 \end{bmatrix} \begin{bmatrix} \hat{\rho}_{i,j+1}^n - \hat{\rho}_{i,j-1}^n \\ \hat{u}_{i,j+1}^n - \hat{u}_{i,j-1}^n \\ \hat{v}_{i,j+1}^n - \hat{v}_{i,j-1}^n \\ \hat{p}_{i,j+1}^n - \hat{p}_{i,j-1}^n \end{bmatrix} \right. \\ &\quad \left. - \begin{bmatrix} 0 & 0 & 0 & \frac{1}{c_i^n} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & c_i^n & 0 \\ 0 & 0 & 0 & c_i^n \end{bmatrix} \begin{bmatrix} \hat{\rho}_{i,j+1}^n - 2\hat{\rho}_{i,j}^n + \hat{\rho}_{i,j-1}^n \\ \hat{u}_{i,j+1}^n - 2\hat{u}_{i,j}^n + \hat{u}_{i,j-1}^n \\ \hat{v}_{i,j+1}^n - 2\hat{v}_{i,j}^n + \hat{v}_{i,j-1}^n \\ \hat{p}_{i,j+1}^n - 2\hat{p}_{i,j}^n + \hat{p}_{i,j-1}^n \end{bmatrix} \right) + o(\delta). \end{aligned} \quad (37)$$

Then the substitution of (19) and (37) into the definition of Godunov method (5) yields

$$\begin{aligned} \rho_{i,j}^{n+1} &= \rho_{i,j}^n - \frac{\Delta t^n}{\Delta x} \left\{ \rho_{i,j}^n u_{i,j}^n - \rho_{i-1,j}^n u_{i-1,j}^n \right\} \\ &\quad - \frac{\Delta t^n}{2\Delta y} \left\{ \rho_i^n (\hat{v}_{i,j+1}^n - \hat{v}_{i,j-1}^n) - \frac{1}{c_i^n} (\hat{p}_{i,j-1}^n - 2\hat{p}_{i,j}^n + \hat{p}_{i,j+1}^n) \right\} + o(\delta), \end{aligned} \quad (38)$$

$$\begin{aligned} \rho_{i,j}^{n+1} u_{i,j}^{n+1} &= \rho_{i,j}^n u_{i,j}^n - \frac{\Delta t^n}{\Delta x} \left\{ \rho_{i,j}^n (u_{i,j}^n)^2 - \rho_{i-1,j}^n (u_{i-1,j}^n)^2 + p_{i,j}^n - p_{i-1,j}^n \right\} \\ &\quad - \frac{\Delta t^n}{2\Delta y} \left\{ \rho_i^n u_i^n (\hat{v}_{i,j+1}^n - \hat{v}_{i,j-1}^n) - \frac{u_i^n}{c_i^n} (\hat{p}_{i,j-1}^n - 2\hat{p}_{i,j}^n + \hat{p}_{i,j+1}^n) \right\} + o(\delta), \end{aligned} \quad (39)$$

$$\begin{aligned} \rho_{i,j}^{n+1} v_{i,j}^{n+1} = & \rho_{i,j}^n v_{i,j}^n - \frac{\Delta t^n}{\Delta x} \{ \rho_{i,j}^n u_{i,j}^n v_{i,j}^n - \rho_{i-1,j}^n u_{i-1,j}^n v_{i-1,j}^n \} \\ & - \frac{\Delta t^n}{2\Delta y} \{ (\hat{p}_{i,j+1}^n - \hat{p}_{i,j-1}^n) - \rho_i^n c_i^n (\hat{v}_{i,j-1}^n - 2\hat{v}_{i,j}^n + \hat{v}_{i,j+1}^n) \} + o(\delta), \end{aligned} \quad (40)$$

$$\begin{aligned} e_{i,j}^{n+1} = & e_{i,j}^n - \frac{\Delta t^n}{\Delta x} \{ u_{i,j}^n (e_{i,j}^n + p_{i,j}^n) - u_{i-1,j}^n (e_{i-1,j}^n + p_{i-1,j}^n) \} \\ & - \frac{\Delta t^n}{2\Delta y} \left\{ \left(\frac{\rho_i^n (u_i^n)^2}{2} + \frac{\gamma p_i^n}{\gamma-1} \right) (\hat{v}_{i,j+1}^n - \hat{v}_{i,j-1}^n) - \left(\frac{(u_i^n)^2}{2c_i^n} + \frac{c_i^n}{\gamma-1} \right) (\hat{p}_{i,j-1}^n - 2\hat{p}_{i,j}^n + \hat{p}_{i,j+1}^n) \right\} + o(\delta). \end{aligned} \quad (41)$$

Multiplying (38) by $-v_{i,j}^n$, adding each of the both sides to those of (40) and deviding the both sides of summation by ρ_i^{n+1} , we obtain

$$\begin{aligned} \hat{v}_{i,j}^{n+1} = & \hat{v}_{i,j}^n - \frac{\Delta t^n}{\Delta x} \frac{\rho_{i-1}^n}{\rho_i^{n+1}} \cdot u_{i-1}^n (\hat{v}_{i,j}^n - \hat{v}_{i-1,j}^n) \\ & - \frac{\Delta t^n}{\Delta y} \left\{ \frac{1}{2\rho_i^{n+1}} (\hat{p}_{i,j+1}^n - \hat{p}_{i,j-1}^n) - \frac{\rho_i^n}{\rho_i^{n+1}} \cdot \frac{c_i^n}{2} (\hat{v}_{i,j-1}^n - 2\hat{v}_{i,j}^n + \hat{v}_{i,j+1}^n) \right\} + o(\delta) \end{aligned} \quad (42)$$

This completes the proof.

5 Numerical Experiments

We verify the formula (15) in theorem 1 by numerical calculation of a progresing planar shock wave.

We assume a flow including a shockwave in xy -plane (two dimension). The shock surface is a line parallel to y -axis, and the velocity vector of flow is parallel to x -axis, *i.e.* the y -component of velocity vector is 0. We simulate the situation by numerical calculation. The condition of numerical calculation is the following. Every physical value is treated in non-dimensional manner.

1. The mesh size (number of finite volumes) is 400 (in x -dirextion) \times 20 (in y -dirextion). The mesh is uniform and $\Delta x = \Delta y$. *i.e.* We extract the rectangular region $[x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$, where $x_{\max} - x_{\min} = 400\Delta x$, $y_{\max} - y_{\min} = 20\Delta y$, and then we divide the region into 400×20 rectangular finite volumes of the same size $\Delta x \times \Delta y$.
2. The temporal increment Δt^n is determined to satisfy the assumption 3. Practically, let the constant C in (14) be 0.7 and then we take the equality of (14) to determine Δt^n .
3. The left state and the right state of shock wave are given as the following table.

	Left	Right
Pressure ρ	5.999	1.0
Velocity u	20.60	1.0
Pressure p	460.9	0.01

The adiabatic constant γ is 1.4.

4. The boundary is treated in the following manner.

- a) Supersonic inflow condition at $\{x_{\min}\} \times [y_{\min}, y_{\max}]$
- b) Supersonic outflow condition at $\{x_{\max}\} \times [y_{\min}, y_{\max}]$
- c) Slipping boundary condition at $[x_{\min}, x_{\max}] \times \{y_{\min}, y_{\max}\}$

5. We use the scheme of Godunov method (5)-(9).

We mention that the situation of numerical calculation satisfies the assumption 1-3 in the previous section.

Godunov method (5)-(9) is of the first order accurate. Therefore the shock that is captured by numerical computation is smeared, *i.e.* includes several intermediate states between the left and right states. The smearing is caused by the scheme's own numerical viscosity. On the other hand we remember the property that the characteristics associated with a shock wave collide with the shock wave from the both sides. The property still works in the numerical calculation, *i.e.* the information of left and right states always tend to propagate toward the shock wave. Therefore the numerical smearing of shock wave does not expand beyond some extent. In fact, after some time steps the profile of shock wave is somehow stable and includes around 20 intermediate states between the left and right states. See figure 1. (Between the two lines, there are 20 nodes.)

As the time is going on, the occurrence and growth of numerical carbuncle is observed. Figure 2 shows the situations at the time $t = 1, 5, 7, 10$. At $t = 1$ the carbuncle is not yet seen, and at $t = 5$ it is still rather small. Then, it grows rapidly and the carbuncle nearly destroys the calculation at $t = 10$. This is never a good numerical result, but it is a good demonstration of numerical carbuncle.

Using the numerical data we now examine to which extent the formula (15) is valid in the real numerical carbuncle. We estimate the left hand side $\hat{v}_{i,j}^{n+1}$ of (15) taking the value of $v_{i,j}^{n+1}$ from the real calculation. We also calculate the right hand side with the values from real calculation, which means that to $\hat{v}_{i,j}^n, \hat{v}_{i-1,j}^n, \hat{v}_{i,j\pm 1}^n, \hat{p}_{i,j+1}^n - \hat{p}_{i,j-1}^n, \rho_{i-1}^n, \rho_i^n$ and ρ_{i+1}^{n+1} , we substitute $v_{i,j}^n, v_{i-1,j}^n, v_{i,j\pm 1}^n, p_{i,j+1}^n - p_{i,j-1}^n, \rho_{i-1,j}^n, \rho_{i,j}^n$ and $\rho_{i,j}^{n+1}$, respectively. Then we calculate the relative error of right hand side (RHS) from the left hand side (LHS);

$$\frac{(\text{RHS}) - (\text{LHS})}{(\text{LHS})}.$$

Tables 1-4 show the relative error, where the integer $\pm M$ ($M \geq 0$) means that the relative error is equal $\pm M\%$ or between $\pm M\%$ and $\pm(M+1)\%$. The tables show only the part of numerically captured shock wave where the carbuncle phenomenon is observed. The left column shows the numbers indicating the position of each node in x -direction.

We observe that the formula follows the amplification of numerical carbuncle rather well. Even at $t = 7$ or $t = 10$, when the numerical carbuncle is rather strong, the coincidence between the left hand side of formula (in some sense, theoretical estimate of the amplification of numerical carbuncle from the time step n to $n+1$) and the right hand side (the real error) is still much better than might be expected from the caotic pictures in Figure 2 by which one would feel that the computation is almost destroyed. We observe some number in the tables are so big as 159 or 51. The violation of assumption 1, especially the fail of linearized estimate (22) for $\bar{G}_{i,j+\frac{1}{2}}^n - \bar{G}_{i,j-\frac{1}{2}}^n$, gives such big numbers. But, generally speaking, we observe that the formula (15) demonstrates the propagation and amplification of numerical carbuncle rather well.

We conclude that the formula (15) in theorem 1 gives a good explanation of the growth of numerical carbuncle until it becomes too big to apply the linear approximation (21) to Riemann problem (8).

6 Concluding Remarks

The discussion in this article gives some theoretical explanation on the growth or amplification of numerical carbuncle. It is interesting that linear approximation in the direction parallel to the shock surface is useful to obtain the explanation. We also mention that the nonlinearity of the compressible Euler equations results in the ratios $\frac{\rho_{i-1}^n}{\rho_i^{n+1}}$ and $\frac{\rho_i^n}{\rho_i^{n+1}}$ of the density when we obtain the formula (5).

The discussion here is still restricted to some specialized case. It may be an interesting problem to extend the analysis to some more general case, while the extension does not seem so direct.

References

- [1] S. K. Godunov. Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics (in Russian). *Mat. Sb. (N.S.)*, 47:251–306, 1959.
- [2] Randall J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhauser-Verlag, Basel, 1990. ISBN 3-7643-2464-3.
- [3] J. Quirk. A contribution to the great Riemann solver debate. *International Journal for Numerical Methods in Fluids*, 18:555–574, 1994.
- [4] J.-Ch. Robinet, J. Gressier, G. Casalis, and J.-M. Moschetta. Shock Wave Instability and Carbuncle Phenomenon: same intrinsic origin? . *J. Fluid Mechanics*, 417:237–263, 2000.

